

WHAT IS CLAIMED IS:

- 1 1. A computer implemented method of identifying and extracting
2 content from HTML formatted web pages, comprising the steps of:
3 selecting a model page, wherein the model page includes a plurality of
4 HTML tags;
5 identifying a first area of interest in the model page;
6 parsing the model page to determine a first string of symbols associated
7 with the plurality of HTML tags, wherein the first area of interest is identified by a first
8 portion of the first string of symbols;
9 retrieving a second web page;
10 parsing the second web page to determine a second string of symbols
11 associated with the HTML tags of the second web page; and
12 comparing the first and second strings to determine whether the second
13 string includes a second portion similar to the first portion of the first string, wherein the
14 second portion corresponds to a second area of interest in the second page.
- 1 2. The method of claim 1, wherein the step of comparing includes
2 applying an approximate pattern matching algorithm to the first and second strings.
- 1 3. The method of claim 1, further comprising the step of storing the
2 first and second areas of interest in a database.
- 1 4. The method of claim 1, further comprising the step of extracting
2 the second area of interest from the second page.
- 1 5. The method of claim 4, further comprising the step of applying a
2 regular expression matching algorithm to the extracted second area of interest.
- 1 6. The method of claim 1, wherein the first and second areas of
2 interest each include two or more distinct sub-areas of the respective page.
- 1 7. The method of claim 1, wherein the step of identifying a first area
2 of interest includes the step of identifying portions of the HTML tags of the model page.

1 8. The method of claim 1, wherein the step of identifying a first area
2 of interest is performed using a manual pointing and selecting device.

1 9. The method of claim 1, wherein the steps of selecting and
2 identifying are performed manually and wherein the remaining steps are performed
3 automatically.

1 10. The method of claim 1, wherein the second web page is retrieved
2 from a remote website over the Internet.

1 11. The method of claim 1, wherein the HTML tags include attributes
2 and attribute values.

1 12. A computer readable medium containing instructions for
2 controlling a computer system to automatically identify and extract desired content from a
3 retrieved HTML formatted web page, by automatically:
4 parsing the HTML code of a manually selected model web page to
5 determine a first string of symbols associated with a first plurality of HTML tags;
6 retrieving a second web page;
7 parsing the HTML code of the second web page to determine a second
8 string of symbols associated with HTML tags of the second page; and
9 comparing the first and second strings to determine whether the second
10 page includes a second plurality of HTML tags substantially matching the first plurality
11 of HTML tags.

1 13. The computer readable medium of claim 12, wherein the first
2 plurality of HTML tags are identified by an operator using a pointing and selection device
3 coupled to the computer system.

1 14. The computer readable medium of claim 12, wherein the second
2 web page is retrieved from a remote website over the Internet.

1 15. The computer readable medium of claim 12, further including
2 instructions for extracting a portion of the second page corresponding to the second
3 plurality of HTML tags.

1 16. The computer readable medium of claim 15, wherein the
2 instructions further control the computer system to store the extracted portion of the
3 second page in a database.

1 17. The computer readable medium of claim 15, further including
2 instructions for controlling the computer system to apply a regular expression matching
3 algorithm to the extracted portion of the second page.

1 18. The computer readable medium of claim 15, wherein the extracted
2 portion of the second page includes two or more distinct sub-areas.

1 19. The computer readable medium of claim 12, wherein the
2 instructions for comparing include instructions for applying an approximate string
3 matching algorithm to the first and second strings.

1 20. The computer readable medium of claim 12, wherein the HTML
2 tags include attributes and attribute values.

1 21. A computer system for identifying and extracting content from
2 HTML formatted web pages, the system comprising:

3 means for retrieving web pages including HTML tags, wherein a model
4 web page is retrieved;

5 means for manually identifying a first area of interest in the model page,
6 wherein the first area of interest corresponds to a first plurality of HTML tags; and

7 a processor including:

8 means for parsing a page, wherein the parsing means parses the
9 model page to determine a first string of symbols associated with the first plurality of
10 HTML tags, and wherein the parsing means thereafter parses an automatically retrieved
11 second web page to determine a second string of symbols associated with the HTML tags
12 of the second web page;

means for comparing the first and second strings to determine whether the second string includes a second portion similar to the first portion of the first string, wherein the second portion corresponds to a second area of interest in the second page; and

means for extracting the second area of interest from the second page.

22. A computer implemented method of identifying and extracting content from web pages formatted using a markup language, comprising the steps of:
selecting a model page, wherein the model page includes a plurality of tokens;
identifying a first area of interest in the model page;
parsing the model page to determine a first string of symbols associated with the plurality of tokens, wherein the first area of interest is identified by a first portion of the first string of symbols;
retrieving a second web page;
parsing the second web page to determine a second string of symbols associated with the tokens of the second web page; and
comparing the first and second strings to determine whether the second string includes a second portion similar to the first portion of the first string, wherein the second portion corresponds to a second area of interest in the second page.

23. The method of claim 22, further comprising the step of extracting the second area of interest from the second page.

24. The method of claim 22, wherein the markup language is selected from the group consisting of HTML, XML, WML, DHTML and HDML.

25. The method of claim 22, wherein the tokens include tag elements and text elements.